IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

First Named Inventor: Pradeep Bahl	Attorney Docket No.: 150789.01
Application No.: 09/714,406	Group Art Unit: 2141
Filed: 11/16/2000	Confirmation Number: 5052
Customer No.: 69316	Examiner: Djenane Bayard
Title: SYSTEM AND METHOD FOR PERFORMING CLIENT-CENTRIC LOAD BALANCING OF MULTIPLE GLOBALLY-DISPERSED SERVERS	

Mail Stop Appeal Brief-Patents
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

APPEAL BRIEF

Sir:

Pursuant to 37 C.F.R. §41.37, Applicant submits this Appeal Brief for the above-mentioned patent application. Accordingly, Applicant appeals to the Board of Patent Appeals and Interferences seeking review of the Examiner's rejections.

CERTIFICATE OF MAILING OR TRANSMISSION
(Under 37 CFR § 1.8(a)) or ELECTRONIC FILING

I hereby certify that this correspondence is being deposited with the United States Postal Service with sufficient postage as first class mail in an envelope addressed to: Commissioner for Patents, P. O. Box 1450, Alexandria, VA 22313-1450 or facsimile transmitted to the U.S. Patent and Trademark Office on the date shown below.

7/20/2009
Date

James Strom

07/24/2009 HVUONG1 00000038 500463 09714406
02 FC:1402 540.00 DA

I. Real Party in Interest.

The real party in interest is Microsoft Corporation, the assignee of all right, title and interest in and to the subject invention.

II. Related Appeals and Interferences.

None.

III. Status of Claims.

Claims 1-15, 20, 23-25, and 27-29 stand rejected and are pending in the Application.

Claims 1-15, 20, 23-25, and 27-29 are appealed.

Claims 16-19, 21, 22, and 26 stand cancelled.

IV. Status of Amendments.

A Final Office Action was mailed on 8/20/2008. No claim amendments have been submitted since the Final Office Action was mailed.

V. Summary of Claimed Subject Matter.

The pending independent claims are claims 1, 10, 12, 20, 23, 28, and 29.

Claim 1:

A system for performing client-centric load balancing of multiple globally-dispersed servers (Fig. 2, servers 200, 202), the servers being accessed by clients (Fig. 2, client 208) connecting through an ISP having a domain name server (DNS-ISP) (Fig. 2, DNS-ISP 206), the servers further having an authoritative domain name server (DNS-

A) (Fig. 2, DNS-A 204) associated therewith and an external domain name server (DNS-B) (Fig. 2, DNS-B 218), the system comprising:

one of a plurality of load balancing domain name servers (DNS-LBs) deployed in a physical proximity from which the actual network latency of the clients to the multiple globally-dispersed servers may be measured (p. 6, lines 3-11), the DNS-LBs having stored therein IP address information of the multiple globally-dispersed servers to be load balanced (p. 17, lines 17-22; p. 18, lines 8-11; p. 22, lines 9-12), the DNS-LBs each sending mapping information to the DNS-B relating the DNS-LB's IP address to an IP address of the DNS-ISP (p. 18, lines 10-13) to which the DNS-LB is in a physical proximity from which the actual network latency of the clients to the globally-dispersed servers may be measured (p. 6, lines 5-9), the DNS-LBs determining performance characteristics of each of the multiple globally-dispersed servers (p. 20, line 17, to p. 21, line 4), a DNS-LB receiving DNS lookup requests sent from its respective physically-proximate clients to the DNS-LB's corresponding DNS-ISP (p. 20, lines 1-16; p. 21, line 25 to p. 22, line 2), the DNS lookup requests comprising respective hostsnames of some of the globally-dispersed servers (p. 19, lines 23-25; p. 3, lines 7-14), the DNS-LB using its measurements of actual network latency from the clients to the globally-dispersed servers (p. 18, lines 1-5; p.5, lines 19-21; p. 6, lines 3-9) to resolve the DNS lookup requests to respective IP addresses of the some of the globally-dispersed servers, where DNS lookup request's hostname can be resolved to multiple of the IP addresses and the DNS-LB returns to the client the IP address that has lower network latency (p. 21, lines 5-9).

Claim 10:

A method of performing client-centric load balancing of multiple globally-dispersed servers (Fig. 2, servers 200, 202), the servers being accessed by clients (Fig. 2, client 208) connecting through an ISP having a domain name server (DNS-ISP)) (Fig. 2, DNS-ISP 206), the servers further having an authoritative domain name server (DNS-A) (Fig. 2, DNS-A 204) associated therewith, the method comprising the steps of:

receiving IP address information from the DNS-A for the servers to be load balanced (p. 17, lines 17-22; p. 18, lines 8-11; p. 22, lines 9-12);

providing the IP address information to a plurality of load balancing domain name servers (DNS-LB) (p. 17, lines 17-22);

receiving mapping information associating DNS-ISP IP address information to IP address information of a DNS-LB (p. 6, lines 18-20; p. 16, line 14, to p. 17, line 2) located in a physical proximity from which the actual network latency from the clients to the globally-dispersed servers is measured by the DNS-LB from a location physically proximate to the ISP's point of presence (col. 6, lines 5-9; p. 22, line 22, to p. 23, line 8; p. 24, lines 3-4); and

referring DNS address inquiries from a DNS-ISP to a physically proximate DNS-LB in accordance with the mapping information (p. 16, line 23, to p. 17, line 2), a DNS address inquiry comprising a hostname corresponding to multiple of the globally-dispersed servers (p. 19, lines 23-25; p. 3, lines 7-14), and wherein the DNS-LB selects one of the multiple servers according to the DNS-LB's determining of client-to-server latency performance and answers the DNS address inquiry by returning the IP address of the selected server (p. 18, lines 1-5; p.5, lines 19-21; p. 6, lines 3-9; p. 21, lines 5-9) .

Claim 12:

A method of performing client-centric load balancing of multiple globally-dispersed servers (Fig. 2, servers 200, 202), the servers being accessed by clients (Fig. 2, client 208) connecting through an ISP having a domain name server (DNS-ISP) (Fig. 2, DNS-ISP 206), the servers further having an authoritative domain name server (DNS-A) associated therewith (Fig. 2, DNS-A 204) , the method comprising the steps of:

obtaining, by a load balancing domain name server (DNS-LB), IP address information for a DNS-ISP (p. 6, lines 3-9; p. 16, lines 8-11), the DNS-LB located in a physical proximity from which the actual network latency of the clients may be measured (p. 16, lines 2-11; p. 6, lines 5-9; p. 18, lines 13-17; p. 19, lines 5-7);

providing a mapping of an IP address of the DNS-LB to the IP address information of the DNS-ISP to an external domain name server (p. 6, lines 18-20; p. 16, line 14, to p. 17, line 2);

receiving IP address information for the servers (p. 17, lines 17-22; p. 18, lines 8-11; p. 22, lines 9-12);

monitoring performance of the servers at the received IP addresses by the DNS-LB transmitting communications to the IP addresses of the servers (p. 7, lines 13-15; p. 20, line 17, to p. 21, line 4);

receiving at the DNS-LB a DNS name query that was sent from one of the clients to the DNS-ISP, the DNS name query querying for an IP address of a hostname that corresponds to multiple of the servers (p. 19, lines 20-20; p. 20, lines 11-16); and

providing at least one IP address for a server in response to the DNS name query by selecting the server, based on the monitoring step, from among the multiple servers

that correspond to the hostname, and by returning the IP address for the selected server (p. 20, lines 17-24).

Claim 20:

A method of performing client-centric load balancing of multiple globally-dispersed content servers for handling content requests from clients (Fig. 2, servers 200, 202), the servers being accessed by clients (Fig. 2, client 208) connecting through an Internet service provider's (ISP's) point of presence (POP), the ISP having a domain name server (DNS-ISP) (Fig. 2, DNS-ISP 206), the servers further having an authoritative domain name server (DNS-A) (Fig. 2, DNS-A 204) associated therewith containing information regarding the IP addresses of the servers (p. 17, lines 17-22; p. 18, lines 8-11; p. 22, lines 9-12), the method comprising the steps of:

deploying a plurality of load balancing domain name servers (DNS-LBs) in a physical proximity from which the actual network latency of the clients connecting to the ISP POPs may be measured (col. 6, lines 5-9; p. 22, line 22, to p. 23, line 8; p. 24, lines 3-4);

communicating IP address information for a plurality of second level domain name servers (DNS-Bs) to the DNS-As to enable the DNS-As to refer name queries to the DNS-Bs (p. 16, line 23, to p. 18, line 7);

providing, by the DNS-LBs to the DNS-B, mapping information associating IP addresses of the DNS-LBs to IP addresses of their corresponding DNS-ISPs to enable the DNS-B to refer name queries from DNS-ISPs to DNS-LBs (p. 6, lines 18-20; p. 16, line 14, to p. 17, line 2) ; and

communicating IP address information of the servers to the DNS-LBs (p. 17, lines 17-22; p. 18, lines 8-11; p. 22, lines 9-12);

monitoring, by the DNS-LBs, performance of the servers (p. 7, lines 13-15; p. 20, line 17, to p. 21, line 4);

receiving at a DNS-LB a DNS name query that was sent from one of the clients to the DNS-ISP, the DNS name query querying for an IP address of a hostname that corresponds to multiple of the servers (p. 19, lines 20-20; p. 20, lines 11-16), the DNS name query having been sent from the client in response to the client starting a service request needing an IP address for the hostname (p. 24, lines 1-2); and

providing, by the DNS-LB, in response to the DNS name query from the DNS-ISP, the IP address of a server by selecting the IP address from among the IP addresses of the multiple of the servers based on the step of monitoring (p. 18, lines 1-5; p.5, lines 19-21; p. 6, lines 3-9; p. 20, lines 17-24; p. 21, lines 5-9).

Claim 23:

A method for load balancing content servers, each of the content servers being associated with a same domain name (p. 3, lines 7-11; p. 19, lines 20-25), the method comprising:

receiving a DNS request to resolve the domain name from an ISP DNS server (p. 16, line 23, to p. 17, line 2);

identifying at least one load balancing server from a group of load balancing servers, the identified load balancing server measuring network latency from the load balancing servers to the content servers (p. 18, lines 1-4; p. 20, lines 17-24; p. 6, lines 3-8; p. 21, lines 12-16; p. 20, lines 17-24); and

sending the IP address of the identified load balancing server to the ISP DNS server (p. 16, line 23, to p. 17, line 2), the identified load balancing server configured to select, based on its measuring of network latency, at least one of the content

servers and to resolve the domain name with an IP address associated with the selected content server (p. 18, lines 1-5; p.5, lines 19-21; p. 6, lines 3-9; p. 21, lines 5-9).

Claim 28:

A method comprising:

receiving at a load-balancing domain name service server (DNS-LB) a DNS lookup request received by and redirected from a domain name service server of an internet service provider (DNS-ISP) (p. 16, line 23, to p. 17, line 2), the request having been sent by a client of the DNS-ISP (p. 20, lines 1-16; p. 21, line 25 to p. 22, line 2), the request containing a hostname corresponding to a plurality of IP addresses of servers serving content (p. 21, lines 5-9; p. 19, lines 23-25; p. 3, lines 7-14), where the request was forwarded by the DNS-ISP when the DNS-ISP: determined that an IP address for the hostname was not cached at the DNS-ISP (p. 19, line 22, to p. 20, line) and obtained the IP address of the DNS-LB by issuing a DNS query for the hostname (p. 20, lines 1-16; p. 20, line 23, to p. 21, line 2);

measuring network latency from the DNS-LB to the servers that correspond to the hostname in the request by repeatedly sending communications from the DNS-LB to the servers (p. 18, lines 1-4; p. 20, lines 17-24);

in response to receiving the redirected DNS lookup request of the client at the DNS-LB (p. 20, lines 11-16), selecting an IP address of one of the servers that correspond to the hostname, where the IP address is selected from among the servers based on the measuring of network latency to the servers (p. 18, lines 1-5; p.5, lines 19-21; p. 6, lines 3-9; p. 21, lines 5-9); and

returning to the client the selected IP address (p. 21, line 1).

Claim 29:

A method performed by a DNS server that provides DNS service for a plurality of clients, the method comprising:

receiving from a client a DNS lookup request requesting an IP address for a server having a hostname specified by the request (p. 19, lines 20-20; p. 20, lines 11-16);

in response to receiving the client DNS lookup request, determining that an IP address for the requested hostname is unavailable on the DNS server (p. 19, line 22, to p. 20, line) and in response issuing a DNS query for the hostname (p. 20, lines 1-16; p. 20, line 23, to p. 21, line 2);

in response to issuing the DNS query for the hostname by the DNS server, receiving a referral to an authoritative DNS server (DNS-A) that corresponds to the hostname (p. 15, lines 14-17; p. 20, lines 1-3), the referral providing an IP address of a domain name service load-balancing server (DNS-LB) and causing the DNS server to query the DNS-LB for an IP address of the hostname in the client request (p. 20, lines 1-16);

in response to querying the DNS-LB, receiving an IP address from the DNS-LB (p. 17, lines 20-24), where the IP address corresponding to the hostname was selected by the DNS-LB based on network measurements obtained by repeated transmissions from the DNS-LB to IP addresses that correspond to the hostname (p. 18, lines 1-4; p. 20, lines 17-24); and

sending the IP address received from the DNS-LB to the client that sent the lookup request for the web server having the hostname specified by the request (p. 21,

line 1).

VI. Grounds of Rejection to be Reviewed on Appeal.

1. The rejection of claims 1-11 under 35 U.S.C. § 103 as obvious over U.S. Patent No. 6,671,259 to He in view of Cisco Distibuted Director, by Delgadillo.

Item 6 of the Final Office Action lists claims 1-15, 20, 23-25, and 27-29 as being obvious over He in view of Delgadillo. However, Delgadillo is referred to only in items 6a- 6k of the Final Action, which reject claims 1-11.

2. The rejection of claims 12-15, 20, 23-25 and 27-29 under 35 U.S.C. § 103 as obvious over He in view of U.S. Patent application No. 2005/0022203 to Zisapel.

VII. Argument.

A. Summary of teachings of He, Delgadillo, and Zisapel.

1. He

1-a. He does not measure latency.

The rejection suggests that He's LB server measure latency. However, the system in He only measures network *load* and server load.

He summarizes the nature of its measurement at col. 4, lines 12-20 (emphasis added):

> In the present invention, the network measurements includes the measurement of load or network traffic experienced by a server. Load refers to the total amount of client requests which the server is servicing or the total amount of operations being performed by a server. Network traffic is the total amount of data packets (traffic on the network) being carried to each of the servers from the client systems, as well as, from each of the servers to the client systems.

Rather than measuring latency, He measures data volume and counts client requests.

1-b. He's load balancing (LB) servers only *acquire* load measures or measure received packets, they do not themselves transmit/communicate to content/web servers for the purpose of measuring latency.

The rejection posits that He's load balancing servers themselves measure network latency. See the Final Action, p. 5, lines 8-12; p. 10, lines 10-14; etc., citing col. 4, lines 5-24, and col. 6, lines 30-67 of He.

Preliminarily, the rejection equates the load measures of He with network

latency. Applicant disagrees, but for discussion only will treat them as equivalent.

He's LB servers either (1) obtain load/latency information from another source, or (2) they measure the volume of pass-through network traffic.

Regarding point (1), He describes the LB server as obtaining the network measurements (col. 4, line 10, 21, and 34) which are *provided* to the LB server ("network measurements are performed on each of the *servers* or gathered from each server using network measurement devices", col. 4, lines 9-11, note "server" and "LB server" are two distinct terms; "LB server 67 examines ... the network measurements provided for the servers 69 and 71", col. 5, lines 62-63; "LB server examines ... the network measurements provided for the servers", col. 12, lines 22-23; "network measurements provided for the servers", col. 11, line 67; "network measurements *received* [not performed] by the load balancing server", col. 16, lines 36 and 52).

Regarding point (2), where He describes the LB server itself performing measurement, it is only measurement of the actual traffic that is *passing through* the LB server (col. 8, lines 45-67). The LB server alternates between a macro-control balancing mode and a micro-control balancing mode. In micro-control mode, "the server monitors the information flow on a packet per packet basis and determines the appropriate server to service each packet", col. 7, lines 55-57. In macro-control mode, the "the LB server examines the amount of client requests the LB server is receiving" (col. 13, lines 48-49). To the extent that the LB server actually measures, it is passive measurement, not measurement to a server for the purpose of measuring latency or the like.

In sum, He's LB server only acquires measurements by either obtaining network load measurements from another source, measuring the number of client requests, or measuring pass-through traffic volume. Thus, in no case does He's LB server actually

communicate to a server (e.g., Web/content server) for the purpose of measuring latency thereto.

1-c. He's servers are located near servers being balanced, not clients.

The rejection acknowledges this aspect of He (Final Action, p. 6, lines 13-14; p. 8, lines 7-11; p. 10, lines 1-3; etc.).

1-d. Improbable to modify He by placing LB servers near clients and away from proximity to servers.

As discussed above, He's LB servers measure traffic packet-by-packet and switch servers on a packet-by-packet basis. He's LB server "monitors the information flow on a packet per packet basis ... the LB server can dynamically change from one server to another quickly and during the same session or connection between the client system and the server" (col. 7, lines 55-60). In macro-control mode, "the LB server ... constantly monitor[s] the information flow on a packet per packet basis and determines the appropriate server to service each packet" (col. 8, lines 10-13). This type of fine-grained control of server traffic would be impossible to perform if the LB server were placed in proximity to clients (e.g., far across the network) and removed from the proximity of the servers that it is balancing. It would also be unlikely that an LB server could measure packet-by-packet network packets to a group of servers without proximity thereto.

2. Delgadillo

2-a. Delgadillo measures latency *from* Agents located near servers *to* clients.

Delgadillo teaches two basic approaches for selecting a server to resolve a DNS

lookup request; the use of IGP and BGP routing table metrics (hop distances), and the use of link latency metrics (page 2, right column, items 1. and 2.). The teachings related to hop distances (IGP/BGP info.) are not relevant to the instant case, which deals with latency. In other words, while the Examiner is correct that Delgadillo tries to resolve to a "close" server, only the latency (round-trip, or "RTT") teachings of Delgadillo are relevant for balancing, because the present claims recite using latency not topological proximities.

Among the five specific measurement techniques mentioned at pages 4-6 of Delgadillo (DRP-External, DRP-Internal, DRP-Server, DRP-MED, and DRP-RTT), *only the DRP-RTT ("round trip times") teachings relate to latency.* DRP-External, DPR-Internal, DRP-Server, and DRP-MED are not based on network latency but rather are based on topological measures such as autonomous system (AS) hop counts, which are significantly different than latency measures, as noted by Delgadillo at page 3, second paragraph. In sum, hop counts (IGP and BGP information) do not reveal latency, which is why Delgadillo teaches a separate technique to measure latency.

Regarding Delgadillo's measurement of *latency*, Delgadillo describes this approach at page 6, right column ("DRP-RTT"). Specifically, Delgadillo states (emphasis added):

> [DRP-RTT] enables the DistributedDirector to include server-to-client (since most data travels from server-to-client) round trip times (RTTs) in traffic redirection decisions. This configuration metric enables the network administrator to optimize server load distribution based on server-to-client link latency, resulting in maximized end-to-end server access performance. When the DRP-RTT metric is configured, the DistributedDirector issues a DRP-RTT query to each DRP Server Agent. Upon receipt of these queries, each DRP Server Agent determines the round trip time (link latency) between itself and the requesting client.

> The Director can then identify the "best" server as that associated with the DRP Server Agent returning the lowest round trip time within a specified tolerance level.

In Delgadillo, an Agent, per instruction from the DistributedDirector, finds network latency by determining the latency "between itself [the Agent] and the requesting client". Furthermore, in the first paragraph of page 3, Delgadillo mentions "DRP Agents are also used to determine the round-trip times (link latencies) between the distributed servers and clients." In Delgadillo, latency is measured by a DRP Agent, and the measure is necessarily *from* the DRP Agent and *to* the client. Were the client performing the measurement, the DistributedDirector would not need to tell the Agent to perform the measurement.

This is consistent with Delgadillo's figures. Figure 1b (page 2) shows "DRP Client-to-Server Round-Trip Time Metrics" are gathered and used. Note that the DRP Agents ("DRP") are located near the "Servers", not near the "Client". Note also that "RTT Measurement" dashed lines are between Agents and *clients*. According to Figure 1b, The DistributedDirector, via its DRP Agents, "[m]easures client-to-DRP server round-trip times" ("DRP Agents are also used to determine the round-trip times (link latencies) between the distributed servers and clients", page 3, first para.). It is a DRP Agent, near a Server, that measures latency to a client, not to a Server. Thus the latency measure is a measure *from* the *Agent* and *to* the client. Each figure in Delgadillo shows the DRP Agents located near Servers, and no client is depicted as having a local DPR Agent.

While Delgadillo does use the terminology "client-to-server" in some places, it uses the term "server-to-client" where describing how latency is measured ("server load distribution based on server-to-client link latency", page 6, column 2). This

ambiguous terminology is resolved by Delgadillo's description of how latency is actually measured, which, as shown above, is from a DRP Agent to a client. Note that Delgadillo at several places refers to its Agents as "DRP *Server* Agents" (page 6, column 2). Clearly, an Agent located with a group of Servers and supports the system by measuring in the direction from Servers/Agent to the Client.

For further understanding, see Step 4 at page 11 of Delgadillo:

> the DistributedDirector issues DRP queries to each DRP server agent configured for the subdomain. After the DRP server agents receive the DRP requests, they [Agents] gather the requested DRP metrics. As previously discussed, there are several DRP metrics including an "external metric" an "internal metric" a "server metric" and an "RTT metric." ... the DRP metrics returned provide distances between the DRP server agents/distributed servers and the client's local DNS.

In sum, there are two notable aspects of Delgadillo's DRP-RTT latency measurement. First, the measurement direction is from servers and toward clients. Second, the measurement endpoint is the client, not the server (Delgadillo's agents don't measure latency to servers).

2-b. Delgadillo's Agents are located near Servers, not near clients.

As shown above, Delgadillo's Agents measure *from* Agents and *to* clients. If Delgadillo's Agents were near the clients, any network latency measurement from Agent to client would obviously be trivial and meaningless for server selection. Furthermore, as mentioned above, Delgadillo's figures show DPR Agents near Servers, and Delgadillo refers to Agents as "server agents", noting that "DRP server agents are typically peers to border routers (BGP speakers) that support the distributed end

servers for which DistributedDirector service distribution is desired." By design, DRP server agents are near "end servers". Furthermore, "DRP server agents must have access to full BGP and IGP routing tables" (p. 3, left column, lines 1-3). Generally, for security, access to sensitive routing information such as BGP and IGP routing tables is kept local. Note that the Agents of Delgadillo actually run on the operating system of a router supporting the servers (p. 3, left column, 3d para., and footnote 1), which both minimizes security issues and further shows the improbability of placing agents near clients as opposed to near the servers being balanced.

Moreover, the routers/Agents in Delgadillo are by design located near Web Servers, because the DistributedDirector product is for WWW *server providers* to address scalability issues ("maintaining high *server* availability, *server* performance ... are key challenges faced by today's *WWW sites*. The DistributedDirector seeks to solve these scalability issues", Abstract, 3d para.). The DistributedDirector product is intended and configured for Web service providers, not clients.

Finally, Delgadillo teaches a "Server Availability" technique at a section titled "Server Availability Parameter Redirects Clients to Available Servers" (page 8, left column). Here, Delgadillo mentions that server availability is tested when the "DistributedDirector attempts to create a TCP connection to each of the distributed servers" (emphasis added). There would be no reason for the DistributedDirector server to check server availability if the DRP Agents were already measuring to the servers. Because DRP Agents use TCP probes (see page 6, right column, middle), if the Agents were sending the TCP probes to the servers, they would already know whether the servers were available, and the DistributedDirector would have no need to itself check availability. For this reason also, it is clear that DRP Agents do not measure latency to the Web servers of Delgadillo.

3. Zisapel

3-a. Zisapel's load balancers (LBs, e.g., LB1 and LB2) are near servers being balanced, not clients.

Zisapel discusses load balancers (LBs) using latency (see, e.g., paras. [0017] and [0026]). Zisapel describes its LBs, at para. [0033] (emphasis added):

> Server farms 10 and12 typically comprise a load balancer 16 and 18 respectively, which may be a dedicated load balancer or a server or router configured to operate as a load balancer, with each of the load balancers being connected to one or more servers 20. Load balancers 16 and 18 are alternatively referred to herein as LB1 and LB2 respectively.

According to the plain language of Zisapel, an LB is part of a server farm and is even connected to the servers. Furthermore, the Figures of Zisapel clearly show each server farm as having a directly connected local LB. No figure of Zisapel shows an LB put in proximity to clients. Nor would Zisapel have an LB near clients and away from servers, because, as shown next, the LBs transmit to a client to measure latency *to the client*.

3-b. Zisapel's LBs measure latency *to the client*, not to the server.

Zisapel's LBs use polling to measure latency (para. [0017]). Paragraph [0018] clearly states that polling is performed by an LB pinging or sending a TCP ACK to the client. Paragraph [0040] states that "[t]o determine comparative network proximity, LBI, LB2, and LB3 preferably each send a polling request 58 to client 26 using known polling mechanisms" (emphasis added). Figure 2C of Zisapel unequivocally shows LB servers transmitting "POLLING REQUEST"s originating at the LBs and following lines

with arrows pointing to and ending at the client. Figure 2D shows the LBs receiving "POLLING RESPONSE"s from the client, not from a server.

For brevity, the preceding sub-sections of section A. will be referred to without reference to section A (i.e.,, "3-b" refers to VII.A.3-b).

B. **The rejection of claims 1 and 10 as obvious over He in view of Delgadillo.**

1. Claim 1.

Claim 1 recites "one of a plurality of load balancing domain name servers (DNS-LBs) deployed in a physical proximity from which the actual network latency of the clients to the multiple globally-dispersed servers may be measured". Claim 1 also recites "the DNS-LB using its measurements of actual network latency from the clients to the globally-dispersed servers to resolve the DNS lookup requests".

Notably, claim 1 recites measurements of network latency from clients *to* the *servers*. In other words, the measure has two aspects. First, there is a direction; from DNS-LBs that are physically proximate to clients, to the servers. Second, there is a measurement endpoint; the globally-dispersed servers (note that the DNS-LB of claim 1 is provided with the addresses of the servers).

Regarding the direction of measurement, as shown above in section 2-a, Delgadillo does not measure latency in a direction from client to server, but rather measures latency from an Agent (located near a server) to a client. The direction of latency measurement is not a distinction without a difference. Latency from A to B in an IP network, for example, can differ from the latency of B to A. It has been well known since the inception IP routing that routing can be asymmetric. See "An Experimental Study of Asymmetric Routing" (Karir and Zhang, 1997-1998, available at terpconnect.umd.edu/~karir/papers/wosbis98.pdf). Although published after the date

of the present invention, the following references also discuss the asymmetric nature of IP routing, a protocol not significantly changed since the time of the present invention: "On Routing Asymmetry in the Internet" (2005, www.cs.ucr.edu/~krish/yhe_gcom05.pdf), "What is asymmetric routing?" (date unknown, my.stonesoft.com/support/document.do?docid=1377), among others. Furthermore, aside from the numerous references that describe the asymmetric (direction-sensitive) nature of IP routing, Applicant has previously made note of this phenomena. See the Amendment filed 5/28/2008, page 11, last paragraph ("It is well known in the field of IP routing that routing is not symmetric. That is, the route from node A to node B can be much different than the route from node B to node A. The latency from a client to a server might involve a local or nearby bottleneck which the server communicating to the client might not experience"). Applicant's prior discussion of the potential significance of the direction of latency measure has not been traversed by the Examiner.

The rejection is in error because Delgadillo measures latency from the direction of servers toward clients (from Agents near servers, to routers near clients). In a similar vein, as also shown in section 2-a, Delgadillo does not measure latency to the server.

Regarding the target of latency measurement, claim 1 clearly recites a DNS-LB measuring latency to the servers. As shown in section 2-a, Delgadillo measures to clients, not servers. As shown in section 1-b, He's LB servers do not measure latency from themselves to servers. In fact, as shown in section 1-a, He doesn't even measure latency (see Final Action, p. 5, lines 8-12, which appears to cite He as teaching latency measurement).

Claim 1 recites that DNS-LBs are "deployed in a physical proximity from which

the actual network latency of the clients to the ... servers may be measured, ... a DNS-LB receiving DNS lookup requests sent from its [the DNS-LB's] respective physically-proximate clients". That is, the DNS-LB is in physical proximity to the client such that actual latencies from client to the servers can be measured. He's servers are not near clients (see section 1-c above). The rejection is in error because, as shown in section 1-d, He cannot be modified to move LB servers away from servers and in proximity to clients without defeating its main purpose. As shown in section 2-b, Delgadillo's Agents (which the rejection compares to DNS-LBs) are near servers, not clients.

2. Claim 10.

i. Claim 10 recites a "*client-centric* load balancing method. More particularly, claim 10 recites "clients connecting through an ISP", and "network latency from the clients to the globally-dispersed servers is measured by the DNS-LB from a location physically proximate to the ISP's point of presence".

The rejection, page 8, lines 8-12, notes that He does not teach this feature. The rejection refers to Delgadillo, stating "The DRP server agents are typically peers to border routers that support the distributed end server for which distribution is desired (See page 2, col. 2)" (Final Office Action, page 8, lines 14-16).

First, Delgadillo has no mention of an "ISP's point of presence". Delgadillo doesn't even mention or suggest a client ISP, nor does He. An ISP point of presence (PoP) is a well-known term of art. A search of "ISP point of presence" on any major Internet search engine reveals numerous definitions and well accepted use of the term in the art of computer networking. For example, and without limitation, an ISP PoP can be a physical location where an ISP has equipment (Cisco ISP Essentials, Greene and Smith, 2002, p. 223).

Second, as shown above in section 2-a, Delgadillo does not measure "network latency from the clients *to the globally-dispersed servers*". In contrast, it measures latency from Agents *to clients*.

Third, as shown in section 2-b, Delgadillo's Agents measure from the proximity of servers, which is not "from a location physically proximate to the [client] ISP's point of presence". Moreover, the Examiner's own statement indicates that the agents in Delgadillo "support the *distributed end server*", which is the opposite of supporting a client or a client ISP's point of presence.

Fourth, Claim 10 recites both "clients connecting through an ISP", and "network latency from the clients to the globally-dispersed servers is measured by the DNS-LB from a location physically proximate to the ISP's point of presence". That is, latency is measured from a location physically proximate to a PoP of the ISP that clients connect through. The latency is from the direction of the client's ISP to the server. As shown above in section 2-b, Delgadillo measures latency from the direction of the server toward the client.

Fifth, claim 10 recites "client-to-server latency performance". The rejection cites only He (p. 8, lines 5-7). As shown in section 1-a, He does not measure latency, it measures traffic volume and the number of client requests.

C. **Rejection of claims 12, 20, 23, 28, and 29 as obvious over He in view of Zisapel.**

1. Claim 12.

Claim 12 recites "clients connecting through an ISP", and "the DNS-LB located in a physical proximity from which the actual network latency of the clients may be measured". The rejection notes that He does not teach this feature, and cites Zisapel

(p. 9, line 17, to p. 10, line 3). However, as shown in section 3-a, Zisapel's LBs are near servers being balanced, not clients. And, as shown in section 3-b, Zisapel's LBs measure latency to the client, not to the server (note claim 12 also recites "monitoring performance of the servers at the received IP addresses by the DNS-LB transmitting communications to the IP addresses of the servers", and "selecting the server, based on the monitoring step").

The rejection states that He teaches the monitoring by the DNS-LB (LB server) transmitting communications to the IP addresses of the servers. However, as shown in section 1-a, He does not measure network latency, and as shown in section 1-b, He's LB server either obtains load data from another source, or it measures pass-through packets and client requests.

2. Claim 20.

Claim 20 recites a method of "client-centric load balancing", including "deploying a plurality of load balancing domain name servers (DNS-LBs) in a physical proximity from which the actual network latency of the clients connecting to the ISP POPs may be measured", and "monitoring, by the DNS-LBs, performance of the servers". As shown in sections 1-b, 1-c, and 3-a, He's and Zisapel's load balancers are in proximity to measure *server* latency. Furthermore, as shown in section 1-d, He's LB servers cannot be modified to be away from servers and in proximity from which *client* latency can be measured.

3. Claim 23.

Claim 23 recites "the identified load balancing server measuring network latency from the load balancing servers to the content servers", and "select[ing a server], based

on its [i.e., load balancing server] measuring of network latency". As shown in section 1-a, He does not measure network latency. As shown in section 1-c, He's LB servers are located near the servers and therefore it would not make sense to use a latency measure from the LB servers to the content servers (client connectivity would not be taken into account). Similarly, as shown in section 3-a, Zisapel's LBs are near (even connected to) the servers being balanced, not clients; LB-to-server latency would be trivial and would be independent of the client. As shown in section 3-b, Zisapel's LBs measure latency to the client, not to the server, as recited in claim 23. Furthermore, as shown in section 1-d, He must have its LB server near the servers being balanced; He cannot be modified to have its LB servers in a position to allow for a meaningful measure of latency *to a server*.

4. Claim 28.

Claim 28 recites "measuring network latency from the DNS-LB to the servers that correspond to the hostname in the request by repeatedly sending communications from the DNS-LB to the servers", and "selecting an IP address of one of the servers ... based on the measuring of network latency to the servers". The rejection notes that He does not teach this feature, and refers to Zisapel ("He et al fails to teach where the IP address was selected by the DNS-LB based on transmissions form [sic] the DNS-LB to the IP address that measure network latency from the DNS-LB to the IP address of the server", Final Action, p. 15, lines 8-13).

However, as shown above in section 3-a, Zisapel's LBs are located near server farms and are connected to the servers, therefore measurement of latency *to servers* would not allow a server to be chosen based on client information. Furthermore, Zisapel's servers measure latency from the LB *to the clients* (as shown in section 3-b),

which is in contrast to claim 28's measurement of latency by sending communications from the load balancer (DNS-LB) to the servers being balanced. As shown earlier, latency measures (via transmission or communication between A and B) can be sensitive to the direction of measurement.

In sum, no reference of record teaches a load balancing DNS server (DNS-LB) measuring network latency from itself (the DNS-LB) *to* the servers being balanced. All of the references disclose load balancers residing and operating at the server end.

5. Claim 29.

Claim 29 recites "a method performed by a DNS server that provides DNS service for a plurality of clients", where a hostname is resolved to an IP address. In particular, "the IP address corresponding to the hostname was selected by the DNS-LB based on network measurements obtained by repeated transmissions from the DNS-LB to IP addresses that correspond to the hostname". In other words, the DNS-LB makes network measurements by transmitting from itself to the IP addresses.

The rejection cites paragraphs [0036] to [0038] of Zisapel. However, as shown above, Zisapel positions LB servers with server clusters, and the LB servers measure latency from the LB servers *to the clients*, not to the servers.

He's LB servers use network measurements, but the measurements are not performed by the LB servers, and they are not measurements obtained by transmitting from the LB server to the server IP addresses. He selects a server based on "network measurements" (col. 1, line 9; col. 4, lines 8, 22-23, 27, 33, and 37; col. 5, line 63; col. 8, lines 37 and 47; etc.). According to He, col. 4, lines 13-20:

> the network measurements includes the measurement of load or network traffic experienced by a server. Load refers to the total amount of client requests which the server is

> servicing or the total amount of operations being performed by a server. Network traffic is the total amount of data packets (traffic on the network) being carried to each of the servers from the client systems, as well as, from each of the servers to the client systems.

He's network measurements clearly refers only to the volume of traffic. This type of information is incapable of being measured by an LB server simply transmitting to the server (e.g., pinging, ending an HTTP request, etc.). This is why He describes the LB server as obtaining the network measurements (col. 4, line 10, 21, and 34) which are *provided* to the LB server ("LB server 67 examines ... the network measurements provided for the servers 69 and 71", col. 5, lines 62-63; "LB server examines ... the network measurements provided for the servers", col. 12, lines 22-23; "network measurements provided for the servers", col. 11, line 67; "network measurements *received* [not performed] by the load balancing server", col. 16, lines 36 and 52).

Where He actually describes the LB server performing a type of measurement, it is only measurement of the actual traffic that is *passing through* the LB server (col. 8, lines 45-67). The LB server alternates between a macro-control mode and a micro-control mode. In micro-control mode, "the server monitors the information flow on a packet per packet basis and determines the appropriate server to service each packet", col. 7, lines 55-57. In macro-control mode, the "the LB server examines the amount of client requests the LB server is receiving" (col. 13, lines 48-49). The only measurements performed by the LB server are the amount of client requests and information flow (volume) passing through.

In sum, He teaches LB servers near servers being balanced. The LB servers receive measurements of volume of network traffic as well as server load. The LB

servers select IP addresses based on this information. The only measurement performed by the LB server itself is monitoring the number of connections it is handling. None of the measurement data mentioned in He is measurement performed by the LB server transmitting from the LB server to the servers being balanced.

He cannot be modified to place LB servers away from servers and near or in physical proximity to clients (or DNS of clients' ISP). He's LB servers switch between a micro control mode and a macro control mode:

> "If the LB server is dynamically configured in micro controlled mode, then the LB server provides the IP number of the LB server to represent a path from the client system to a server with the LB server acting as the gatekeeper in step 40" (col. 7, lines 37-40);
>
> "Micro-control mode places the LB server in greater control of the flow of information from the client systems to the servers. If the LB server is dynamically configured to be in micro-control mode, the [LB] server monitors the information flow on a packet 55 per packet basis and determines the appropriate server to service each packet" (col. 7, lines 51-58).

Because LB servers of He monitor the packets of content servers, it is necessary for them to be near the servers. The He LB servers would not operate as intended if they were modified to be proximate to clients rather than servers. Furthermore, they would not be able to control the flow of information. In sum, the LB servers cooperate closely with the content servers, receiving load measurements, handling client traffic for the server, acting as a conduit to dynamically changing servers, etc. (see col. 2, line 23; and col. 6 lines 7-10).

Respectfully submitted,

Microsoft Corporation

Date: 7/20/2009

By: /James T. Strom/

James T. Strom, Reg. No.: 48,702

Attorney for Applicants

Direct telephone 425-939-0781

CERTIFICATE OF MAILING OR TRANSMISSION
(Under 37 CFR § 1.8(a)) or ELECTRONIC FILING

I hereby certify that this correspondence is being deposited with the United States Postal Service with sufficient postage as first class mail in an envelope addressed to: Commissioner for Patents, P. O. Box 1450, Alexandria, VA 22313-1450 or facsimile transmitted to the U.S. Patent and Trademark Office on the date shown below.

7/20/2009
Date

James Strom

VIII. Claims Appendix.

1. (Previously Presented) A system for performing client-centric load balancing of multiple globally-dispersed servers, the servers being accessed by clients connecting through an ISP having a domain name server (DNS-ISP), the servers further having an authoritative domain name server (DNS-A) associated therewith and an external domain name server (DNS-B), the system comprising:

one of a plurality of load balancing domain name servers (DNS-LBs) deployed in a physical proximity from which the actual network latency of the clients to the multiple globally-dispersed servers may be measured, the DNS-LBs having stored therein IP address information of the multiple globally-dispersed servers to be load balanced, the DNS-LBs each sending mapping information to the DNS-B relating the DNS-LB's IP address to an IP address of the DNS-ISP to which the DNS-LB is in a physical proximity from which the actual network latency of the clients to the globally-dispersed servers may be measured, the DNS-LBs determining performance characteristics of each of the multiple globally-dispersed servers, a DNS-LB receiving DNS lookup requests sent from its respective physically-proximate clients to the DNS-LB's corresponding DNS-ISP, the DNS lookup requests comprising respective hostsnames of some of the globally-dispersed servers, the DNS-LB using its measurements of actual network latency from the clients to the globally-dispersed servers to resolve the DNS lookup requests to respective IP addresses of the some of the globally-dispersed servers, where DNS lookup request's hostname can be resolved to multiple of the IP addresses and the DNS-LB returns to the client the IP address that has lower network latency.

2. (Original) The system of claim 1, wherein the DNS-B stores the mapping information for the plurality of DNS-LBs to forward IP address queries to one of the DNS-LBs closest to the DNS-ISP from which the IP address query originated, and wherein the DNS-LB closest to the DNS-ISP returns the IP address to the DNS-ISP of the server having the best performance characteristics.

3. (Original) The system of claim 1, wherein the DNS-B stores the mapping information for the plurality of DNS-LBs to forward IP address queries to one of the DNS-LBs closest to the DNS-ISP from which the IP address query originated, and wherein the DNS-LB closest to the DNS-ISP returns the IP address of the DNS-LB to the DNS-ISP.

4. (Original) The system of claim 1, wherein the DNS-B provides its IP address information to the DNS-A to enable the DNS-A to forward IP address queries to the DNS-B.

5. (Original) The system of claim 4, wherein the DNS-B receives IP address information from the DNS-A for the servers to be load balanced.

6. (Original) The system of claim 1, wherein the DNS-LB is a client of the DNS-ISP.

7. (Original) The system of claim 1, further comprising a DNS-B deployed on each Internet backbone, and wherein each DNS-B contains the mapping information for all of the DNS-LBs stored therein.

8. (Original) The system of claim 1, wherein the DNS-LB transmits updated mapping information upon a change of an IP address of the DNS-ISP.

9. (Original) The system of claim 1, wherein each of the DNS-LBs transmit performance information of the servers to the DNS-B, and wherein the DNS-B utilizes the mapping information to determine the proper DNS-LB performance information to utilize to select the IP address of the server having the best performance characteristics to return to the DNS-ISP from which an IP address query originated.

10. (Previously Presented) A method of performing client-centric load balancing of multiple globally-dispersed servers, the servers being accessed by clients connecting through an ISP having a domain name server (DNS-ISP), the servers further having an authoritative domain name server (DNS-A) associated therewith, the method comprising the steps of:

receiving IP address information from the DNS-A for the servers to be load balanced;

providing the IP address information to a plurality of load balancing domain name servers (DNS-LB);

receiving mapping information associating DNS-ISP IP address information to IP address information of a DNS-LB located in a physical proximity from which the actual network latency from the clients to the globally-dispersed servers is measured by the DNS-LB from a location physically proximate to the ISP's point of presence; and

referring DNS address inquiries from a DNS-ISP to a physically proximate DNS-LB in accordance with the mapping information, a DNS address inquiry comprising a

hostname corresponding to multiple of the globally-dispersed servers, and wherein the DNS-LB selects one of the multiple servers according to the DNS-LB's determining of client-to-server latency performance and answers the DNS address inquiry by returning the IP address of the selected server.

11. (Original) A computer-readable medium having computer executable-instructions for performing the steps of claim 10.

12. (Previously Presented) A method of performing client-centric load balancing of multiple globally-dispersed servers, the servers being accessed by clients connecting through an ISP having a domain name server (DNS-ISP), the servers further having an authoritative domain name server (DNS-A) associated therewith, the method comprising the steps of:

obtaining, by a load balancing domain name server (DNS-LB), IP address information for a DNS-ISP, the DNS-LB located in a physical proximity from which the actual network latency of the clients may be measured;

providing a mapping of an IP address of the DNS-LB to the IP address information of the DNS-ISP to an external domain name server;

receiving IP address information for the servers;

monitoring performance of the servers at the received IP addresses by the DNS-LB transmitting communications to the IP addresses of the servers;

receiving at the DNS-LB a DNS name query that was sent from one of the clients to the DNS-ISP, the DNS name query querying for an IP address of a hostname that corresponds to multiple of the servers; and

providing at least one IP address for a server in response to the DNS name query

by selecting the server, based on the monitoring step, from among the multiple servers that correspond to the hostname, and by returning the IP address for the selected server.

13. (Original) The method of claim 12, further comprising the steps of:

detecting a change in the DNS-ISP IP address; and

updating the mapping of the IP address of the DNS-LB to the IP address information of the DNS-ISP to the external domain name server.

14. (Original) The method of claim 12, further comprising the steps of

receiving selection criteria for the selection of an IP address;

receiving a name query from the DNS-ISP; and

wherein the step of providing at least one IP address for a server in response to a name query selected based on the monitoring step further comprises the step of providing at least one IP address for a server in response to a name query selected based on the monitoring step and on the selection criteria.

15. (Original) A computer-readable medium having computer-executable instructions for performing the steps of claim 12.

16-19. (Canceled).

20. (Previously Presented) A method of performing client-centric load balancing of multiple globally-dispersed content servers for handling content requests from clients, the servers being accessed by clients connecting through an Internet service

provider's (ISP's) point of presence (POP), the ISP having a domain name server (DNS-ISP), the servers further having an authoritative domain name server (DNS-A) associated therewith containing information regarding the IP addresses of the servers, the method comprising the steps of:

deploying a plurality of load balancing domain name servers (DNS-LBs) in a physical proximity from which the actual network latency of the clients connecting to the ISP POPs may be measured;

communicating IP address information for a plurality of second level domain name servers (DNS-Bs) to the DNS-As to enable the DNS-As to refer name queries to the DNS-Bs;

providing, by the DNS-LBs to the DNS-B, mapping information associating IP addresses of the DNS-LBs to IP addresses of their corresponding DNS-ISPs to enable the DNS-B to refer name queries from DNS-ISPs to DNS-LBs; and

communicating IP address information of the servers to the DNS-LBs;

monitoring, by the DNS-LBs, performance of the servers;

receiving at a DNS-LB a DNS name query that was sent from one of the clients to the DNS-ISP, the DNS name query querying for an IP address of a hostname that corresponds to multiple of the servers, the DNS name query having been sent from the client in response to the client starting a service request needing an IP address for the hostname; and

providing, by the DNS-LB, in response to the DNS name query from the DNS-ISP, the IP address of a server by selecting the IP address from among the IP addresses of the multiple of the servers based on the step of monitoring.

21. (Cancelled)

22. (Cancelled)

23. (Previously Presented) A method for load balancing content servers, each of the content servers being associated with a same domain name, the method comprising:

receiving a DNS request to resolve the domain name from an ISP DNS server;

identifying at least one load balancing server from a group of load balancing servers, the identified load balancing server measuring network latency from the load balancing servers to the content servers; and

sending the IP address of the identified load balancing server to the ISP DNS server, the identified load balancing server configured to select, based on its measuring of network latency, at least one of the content servers and to resolve the domain name with an IP address associated with the selected content server.

24. (Previously Presented) The method as recited in claim 23, wherein the certain characteristics include load level, availability, network latency, or network cost.

25. (Previously Presented) The method as recited in claim 23, wherein the identified load balancing server is situated closest to the ISP DNS server among the group of load balancing servers.

26. (Cancelled)

27. (Previously Presented) The system as recited in claim 26, wherein the

certain characteristics include load level, availability, network latency, or network cost.

28. (Previously Presented) A method comprising:

receiving at a load-balancing domain name service server (DNS-LB) a DNS lookup request received by and redirected from a domain name service server of an internet service provider (DNS-ISP), the request having been sent by a client of the DNS-ISP, the request containing a hostname corresponding to a plurality of IP addresses of servers serving content, where the request was forwarded by the DNS-ISP when the DNS-ISP: determined that an IP address for the hostname was not cached at the DNS-ISP and obtained the IP address of the DNS-LB by issuing a DNS query for the hostname;

measuring network latency from the DNS-LB to the servers that correspond to the hostname in the request by repeatedly sending communications from the DNS-LB to the servers;

in response to receiving the redirected DNS lookup request of the client at the DNS-LB, selecting an IP address of one of the servers that correspond to the hostname, where the IP address is selected from among the servers based on the measuring of network latency to the servers; and

returning to the client the selected IP address.

29. (Previously Presented) A method performed by a DNS server that provides DNS service for a plurality of clients, the method comprising:

receiving from a client a DNS lookup request requesting an IP address for a server having a hostname specified by the request;

in response to receiving the client DNS lookup request, determining that an IP

address for the requested hostname is unavailable on the DNS server and in response issuing a DNS query for the hostname;

in response to issuing the DNS query for the hostname by the DNS server, receiving a referral to an authoritative DNS server (DNS-A) that corresponds to the hostname, the referral providing an IP address of a domain name service load-balancing server (DNS-LB) and causing the DNS server to query the DNS-LB for an IP address of the hostname in the client request;

in response to querying the DNS-LB, receiving an IP address from the DNS-LB, where the IP address corresponding to the hostname was selected by the DNS-LB based on network measurements obtained by repeated transmissions from the DNS-LB to IP addresses that correspond to the hostname; and

sending the IP address received from the DNS-LB to the client that sent the lookup request for the web server having the hostname specified by the request.

IX. EVIDENCE APPENDIX.

None.

X. RELATED PROCEEDINGS APPENDIX.

None.